Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 7.1 (2025) 1–26
HTTPS://DOI.ORG/10.5117/CCR2025.1.13.LIND

Grounding the Comparative Turn in Communications: A Framework for Validating Multilingual Computational Text Analysis

Fabienne Lind
Department of Communication Science, University of Vienna, Austria

Martijn Schoonvelde
Faculty of Arts, University of Groningen, The Netherlands

Christian Baden
Department of Communication and Journalism, The Hebrew University of Jerusalem, Israel

Alona O. Dolinsky
Department of Communication Science, Vrije Universiteit Amsterdam, The Netherlands

Christian Pipal
Department of Communication and Media Research, University of Zurich, Switzerland

Mariken A.C.G. van der Velden
Department of Communication Science, Vrije Universiteit Amsterdam, The Netherlands

Abstract
Following the progressing internationalisation of social science research and the computational turn in the field, researchers are increasingly adopting computational text analysis (CTA) methods to compare textual data across multiple cases and languages. In these settings, it is not only the mapping between construct and measures that requires validation, but also the equivalence of this mapping across languages and cases. However, although the validation requirements in multilingual analyses exceed those in monolingual studies, current research shows that validation is often insufficiently and inconsistently addressed in comparative multilingual CTA. To support more robust comparative research, this article presents a framework for validating findings obtained from multilingual textual data. The framework

outlines validation strategies for four key stages of a typical multilingual CTA workflow: corpus, input data, process, and output. It directly tackles the challenge of approaching equivalence across contexts and languages in these stages and moves beyond the common practice of identifying problems only at the final stage of research.

Keywords: cross-lingual, comparative research, text as data, computational text analysis, validation framework, internationalisation

1 Introduction

Over the past decade, the internationalisation of social science research has finally been picking up speed (e.g., Henriksen, 2016; Scharkow & Trepte, 2023). As scholars engage more globally, there is growing recognition of the need to confront ongoing power imbalances in the field (Demeter, 2019), prompting efforts to broaden the scope of research beyond the dominance of WEIRD (Western, Educated, Industrialised, Rich, and Democratic) countries (Henrich et al., 2010). While comparative survey research has embraced this global shift, textual research has lagged, partly due to the cultural and linguistic complexities embedded in texts, which makes comparisons across cases[1] more complex (e.g., Rössler, 2012). Only in recent years have rapid advances in computational methods increasingly allowed researchers to analyse textual data not only at scale, but also in a multilingual and comparative manner.

Multilingual computational text analysis (CTA) is an emerging and fast-evolving field. Scholars apply it to study climate change coverage across Germany, India, South Africa, and the United States (Wozniak et al., 2021), foreign policy reporting in over 100 countries (Baum & Zhukov, 2019), or the salience of the EU in national legislatures (Rauh & De Wilde, 2018). Methodological innovations have made such studies increasingly feasible. These range from multilingual dictionaries (e.g., Baden et al., 2018; Proksch et al., 2019), multilingual supervised machine learning (e.g., Lind et al., 2021), multilingual topic modeling (e.g., Chan et al., 2020; Lind et al., 2022; Lucas et al., 2015; Reber, 2019), multilingual representation models based on transformers (e.g., De Vries, 2021; Laurer, 2023; Licht, 2023), and, most recently, to prompt-based multilingual generative models (e.g., Rathje et al., 2024).

Nevertheless, despite these developments, a persistent challenge remains: there is little consensus on how to validate multilingual CTA for

[1] Cases encompass a spectrum of macro-level units, ranging from countries and regions to markets and beyond.

comparative research. Most validation discussions remain rooted in single-language contexts and provide limited guidance for cross-lingual applications (e.g., Birkenmaier et al., 2024; Grimmer et al., 2022; Song et al., 2020; van Atteveldt & Peng, 2018). Even in single-language contexts, recent reviews find that validation is inconsistently applied and often underreported; for instance, only 54% of CTA articles in top communication journals include any form of validation (Stecker et al., 2024). As Birkenmaier et al. (2024) note, the lack of coherent standards and field-specific guidance for CTA has led to fragmented and ad hoc validation practices, threatening the credibility and comparability of findings.

In multilingual settings, the need for rigorous validation is especially pressing. Recent research shows strong demand for improved validation practices in these contexts. A survey of authors of published CTA studies (Dolinsky et al., 2024) found that those working with multiple languages expressed significantly more concern about the validity of their findings than those focusing on English or a single other language. However, this demand has not translated into widespread implementation of validation strategies. A review of published social science literature (Baden, Dolinsky, et al., 2022)[2] shows that validation in multilingual CTA is neither sufficiently nor consistently addressed.

In this article, we address this gap by proposing a framework for validating multilingual measurement in CTA for comparative research. We argue that to ensure comparability of textual measures across languages and cases, equivalence must be validated at multiple stages of the research process. Our framework identifies four key stages for validation: corpus, input data, process, and output. These stages reflect a typical CTA workflow, offering a structured and actionable approach to validation. The chronological structure helps guide researchers through validation from data collection to final measurement.

Crucially, each stage raises distinct challenges for establishing comparability. For example, without a corpus appropriate to the research objective, later validation steps may lose their meaning. Similarly, problems uncovered during output validation often stem from earlier stages, such as biased input selection or inadequate processing. A key motivation for our four-stage framework is that the need for validation far exceeds demonstrating performance relative to a given benchmark of the final measures, and that corpus, input data, and process constitute key considerations with far-reaching

[2]It builds on a review of practises reported in the quantitative text-based social science literature that was developed as part of the text analysis infrastructure project OPTED (https://opted.eu/) spanning multiple countries and languages

implications for measurement validity.

This perspective also highlights how multilingual comparative text analysis introduces challenges that monolingual, single-context studies largely avoid. Whereas a researcher working in one language 'only' needs to ensure that the measures obtained meaningfully map onto their construct of interest, researchers working in multiple languages and with numerous cases need to provide evidence that this link between the conceptual and empirical realms is equivalent across languages and cases. Only when such equivalence is established can substantial comparisons of cases be deemed meaningful.

This framework is especially timely, as researchers increasingly rely on pre-trained large language models (LLMs), which often lack transparency regarding their training data or language coverage, and tend to perform unevenly across languages due to resource imbalances and structural biases (Bender et al., 2021; Mate et al., 2023). Without systematic validation, such tools risk introducing hidden biases or misinterpreting cross-lingual patterns as substantive findings.

To illustrate how the framework can be applied in practice, we use a running example drawn from the computational detection of incivility. Specifically, we build on the premise of replicating Muddiman and Stroud's (2017) classic study on incivility in online news comment sections in a multilingual, internationally comparative setting. This example is particularly well-suited for our purposes because incivility is expressed in highly variable linguistic and cultural forms. By drawing on this example throughout, we aim to concretely illustrate how language- and context-sensitive validation can be addressed through specific considerations and evaluations at each stage of a typical multilingual CTA pipeline.

While specific validation strategies will necessarily depend on the setting and purpose of each study, the framework is sufficiently general to apply to a wide range of analytical approaches, including dictionary-based models, topic modeling, and transformer-based architectures. It offers guidance to researchers seeking to draw meaningful measurements from multilingual text data in comparative research projects across diverse methodological settings. Developing widely accepted standards for validating the use of CTA methods is crucial for building trust in their application to multilingual comparative research. It represents a necessary prerequisite for their broader adoption in an increasingly global research community.

2 Scope conditions, key terms, and goals

Our validation framework is designed for research that aims for measures that can be validly compared across languages and cases, and thus for comparative research designs (Volk, 2022). As such, it supports the rapidly expanding community of researchers who study text beyond the monolingual English default (Baden, Pipal, et al., 2022). To ensure comparability, the framework follows an etic approach at the level of construct definition, starting from a universal definition of the concept, assumed to involve the same theoretical dimensions across contexts (Esser & Vliegenthart, 2017). While it is possible to compare constructs defined emically (i.e., in context-specific ways), such comparisons are typically qualitative or interpretive. By focusing on etic definitions, our framework ensures that cross-national and cross-linguistic comparisons can be made using shared, conceptually equivalent criteria (Wirth & Kolb, 2004). For example, a cross-national extension of Muddiman and Stroud's (2017) study might adopt their modular definition of incivility, comprising name-calling, profanity, racism, and threats, as a shared baseline. This reflects an etic strategy: maintaining conceptual equivalence across settings. The next step would be to determine how to measure this construct in each of the cases and languages under study.

We take a broad view of measurement validity, referring to the extent to which a given CTA method accurately captures the intended meanings in a text (Grimmer et al., 2022). In multi-case, multi-language research, achieving validity also requires careful attention to equivalence and comparability. When analysing an etic concept across different cases and languages, it is essential that the measures consistently reflect the same underlying construct. This alignment, referred to as "equivalent mapping", is a central goal in comparative research. Following (Volk, 2022, p. 80), we define equivalence as the condition in which "two objects, structures, or categories have an equal value or the same function".

Ensuring equivalence does not mean enforcing a uniform measurement strategy across all contexts. For example, in a comparative study of uncivil discourse, researchers must account for how certain expressions function differently across contexts. A term like "gay", which may be neutral or reclaimed in some settings, can still carry a strongly derogatory connotation in others. Similarly, culturally specific slurs such as the German "grünlinksversifft" ("green-left-filthy") have no direct equivalent in the U.S. political discourse, illustrating how seemingly analogous categories can diverge sharply in language and meaning. The key is that measures must reflect meaningful case differences relevant to the research question. At the same time,

they must avoid distortions caused by unnoticed inconsistencies in measurement. A robust validation process helps distinguish true cross-case differences from those arising due to methodological artifacts (He & van de Vijver, 2012). Both validation goals—equivalent mapping and measure comparability—are summarised in Figure 1.

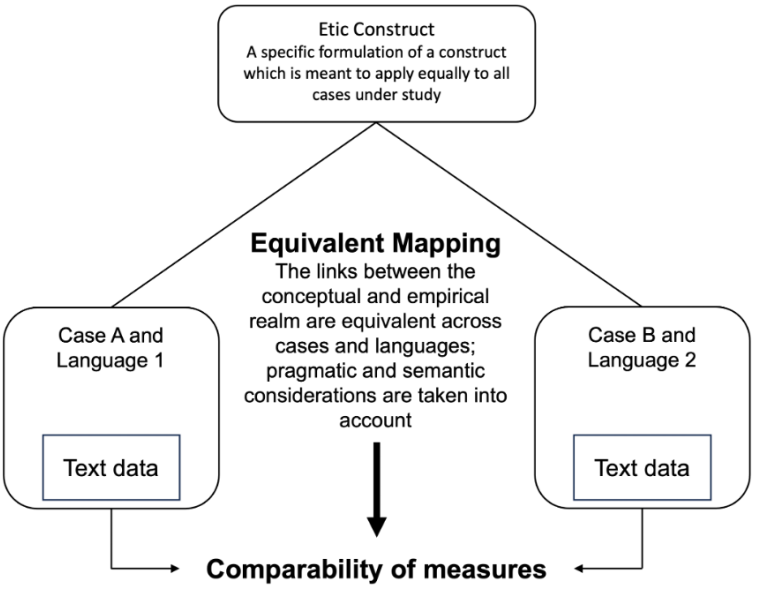

Figure 1: Measurement and validity objectives for comparative research designs. Note. This simplified illustration assumes only two cases, each of which is linked to exactly one language; actual cases may be considerably more complex.

Two linguistic considerations are crucial for establishing equivalence in multilingual comparative designs: semantic comparability, which concerns the literal meaning of language and the relationship between signs and their referents, and pragmatic comparability, which relates to the social meaning of language and how it is interpreted within specific contexts (Hovy & Prabhumoye, 2021; Licht & Lind, 2023).

A key challenge for semantic comparability concerns the extraction of comparable meanings from text data despite significant variation in how they are encoded in each language, through dimensions like script (e.g., Latin, Cyrillic, Hebrew), morphology (word formation), and syntax (see Shababo et al., 2023, for sentence structure). In essence, expressions in different languages do not align, impeding direct numerical comparison. What is more, comparable meanings are often expressed differently across

languages, and the associated connotations or conceptual boundaries can vary as well. For example, a single word in one language may correspond to multiple distinct concepts in another (e.g., consider Hebrew: אויב העם, German: "Volksfeind", but English: "enemy of the people"). Moreover, many words lack straightforward translations altogether (Sigismondi, 2018). In addition, it is possible that some information that is consequential for an intended analysis, such as information on the grammatical case, gender, or tense of expressed contents, is available in some languages, but not in others. As a result, equivalent measurement strategies may need to operate with very different depths of information, potentially distorting results.

The second consideration concerns pragmatic comparability. From a social science perspective, language can rarely be studied without taking into account the context in which language is produced (see also Hovy & Prabhumoye, 2021). In comparative research designs, it is important to consider the divergent meanings and connotations of semantically similar expressions, owing to the influence of historical, political, and social traditions on language use. For example, the label "socialism" may evoke a legacy of authoritarianism in parts of Eastern Europe. At the same time, in conservative U.S. political discourse, it often functions as a generalized slur against progressive policies. For a valid comparative analysis, it is thus crucial to ensure that measured contents are valid not only in terms of their semantic meaning, but also in terms of how equivalent meanings are commonly expressed in different contexts.

3 A framework for validating multilingual textual analysis

Having clarified the challenges of semantic and pragmatic equivalence, we now turn to the framework itself. It provides a structured approach for validating multilingual CTA, with a particular focus on approaching equivalence across languages and contexts. The framework is designed to guide researchers through key decision points in typical CTA workflows, offering validation strategies.

Validation strategies refer to the methods researchers use to ensure that recorded scores accurately reflect the intended concepts (Adcock & Collier, 2001). These strategies may involve reflecting on methodological choices and their alignment with theoretical constructs, as well as empirically testing how well the chosen procedures perform.

The framework (summarised in Figure 2) organizes these strategies across four stages that align with a typical CTA workflow: (1) corpus, (2)

input data, (3) process, and (4) output.

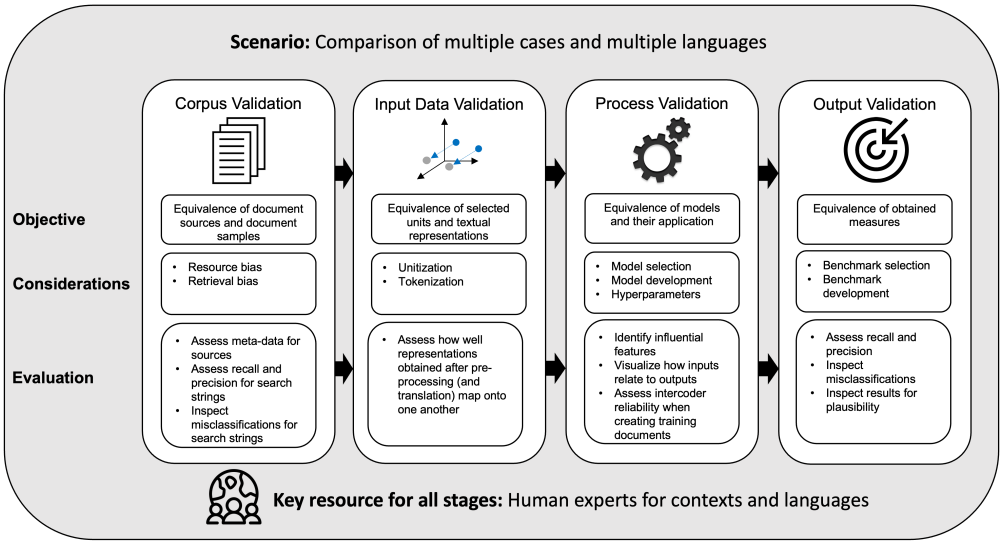

Figure 2: A validation framework for multilingual comparative research designs.

The corpus stage concerns the valid selection of equivalent document sources and samples across multiple cases and languages. The input data stage addresses how these documents should be unitized and quantitatively represented for subsequent text analysis procedures. The process stage involves the application of models, while the output stage focuses on evaluating the resulting measures. These stages follow a largely sequential structure, where each step builds upon the outcomes of the previous one. Problems not identified early can compound in later stages, so each stage should be approached carefully and deliberately to ensure that only validated results progress.

The corpus stage holds a foundational role in the validation framework because errors introduced here are difficult and costly to correct downstream. Document sources may not be easily recollected or reproduced, and platform APIs or repositories often change over time. That said, thorough validation at any stage can never mitigate or replace attention at another stage, as fatal validity and comparability issues may arise independently at every stage. Hence, there is no inherent hierarchy. Each stage is essential, though issues at earlier stages are typically more difficult to remedy later. Problems should be addressed or identified where they happen.

Alternative organizing principles for validation exist, such as psycho-

metric categories (e.g., face, content, criterion, discriminant validity; see Adcock & Collier, 2001; Cronbach & Meehl, 1955) or linguistic levels (e.g., morphological, syntactic, semantic, pragmatic; see O'Grady et al., 2017). Our workflow-based structure integrates these perspectives where they most directly affect research decisions: for example, criterion and face validity during output validation; content validity during input data and process stages; and linguistic challenges as well as pragmatic challenges across the pipeline stages. This chronological structure guides researchers through validation from data collection to final measurement, highlighting key equivalence challenges at each step.

The framework is informed by a systematic review of 854 quantitative text analysis studies published between January 2016 and September 2020 in the top 20 Web of Science journals in communication, political science, sociology, and psychology (Baden, Dolinsky, et al., 2022). We manually coded these studies for reported validation strategies. Only about 45% justified the comparability of document samples across languages and cases. While preprocessing steps were frequently documented, only 5% discuss or show their suitability for achieving comparable multilingual representations. Just 10% reported validation of modeling choices, and only 30% compared outputs with convergent or divergent measures. These findings underscore the lack of standardized validation practices and motivate our contribution.

In what follows, we present the four stages of the framework in greater detail, highlighting key objectives, considerations, and evaluation techniques for each.

3.1 Corpus Validation

3.1.1 Objective

In all comparative research, the validity of findings depends on the equivalence of the materials being compared (Palicki et al., 2023; Rinke et al., 2022). In a text as data project, this means that a researcher must first select equivalent document sources. While not feasible for all research projects, given that some studies rely on "found data" (Salganik, 2017, pp. 82-83) with unknown generative processes, materials need to be sufficiently equivalent in relation to the research question at hand.[3] This guarantees that detected differences are not mere byproducts of subtle differences in the selection of documents. During the process of selecting document sources it is vital to ensure the equivalence of the source sampling procedure across cases, and

[3]Random samples and population-level analyses are unaffected by this step.

make note of any remaining discrepancies (e.g., when specific corpora, for which no suitable equivalent exists in other cases, are included for substantive reasons) so that they can be addressed later in the analysis.

When equivalent document sources have been identified, in many cases, corpus creation furthermore requires the selection of equivalent document samples. This can involve narrowing inclusion criteria to specific relevant document populations (usually, using keywords) or creating stratified samples that capture key dimensions of variability in the population under study. Whenever the cases under study also involve different languages, researchers need to additionally ensure that their retrieval strategies yield equivalent document samples (e.g., translated keywords often find or miss different subsets of relevant debates, owing to divergent connotations and language-specific expressions).[4] To achieve this goal, researchers need to reflect on potential sources of sampling bias and empirically evaluate whether sample equivalence has been achieved.

3.1.2 Considerations

Suppose we wanted to replicate Muddiman and Stroud's (2017) study of news user commentary in the New York Times (NYT) across multiple languages. In such a scenario, a first consideration concerns resource bias and retrieval bias, as recommended by (Grimmer et al., 2022).

Resource bias arises when the availability of documents differs systematically across languages and cases. For their study, Muddiman and Stroud (2017) were able to access all comments posted in response to NYT coverage since the inauguration of its commentary feature, including also comments that were subsequently moderated or deleted. Besides the obvious challenges of determining which media outlet maps well onto the NYT, other media outlets may have opened commentary sections later, selectively, or failed to keep long-term records, and very few, if any, will grant comparable access to moderated content. In some countries, media primarily rely on social media platforms to enable user commentary, subjecting the posting, moderation, and archiving of content to external platform policies and norms. In a similar vein, analyses of X (formerly Twitter) content need to account for the platform's very different roles and adoption rates across contexts, considering whether comparable political discourse might occur on various social networks instead. Media outlets, platforms, and other

[4]When researchers select documents using CTA (i.e., keyword-based methods), this is basically a validation task in itself. It is considered best practice to apply the input, process, and output stages of the validation framework to the document selection step as well.

public forums can differ between countries concerning their policies for moderating or surpressing uncivil content, adding systematic distortions.

Retrieval bias occurs when the sampling strategy for identifying relevant documents works differently between cases in ways that introduce bias into subsequent measurements. For example, a search string for identifying swearwords may have high precision and recall in one language, but perform considerably worse when translated into another language (e.g., Russian, profanity is often rooted in sexual language; in German, it tends to use fecal metaphors; and in the U.S., profanity is frequently abbreviated ["bs"] or blanked ["f***"]).[5] In most cases, it cannot be assumed that identical sampling strategies, applied across cases and languages, will reliably generate equivalent samples.

3.1.3 Evaluation

When evaluating equivalence among language-specific documents, reliance on domain expertise or involvement of case experts becomes essential (Palicki et al., 2023; Rinke et al., 2022). Often, collecting equivalent samples across languages requires comparative reasoning. For instance, in cases where text sources are media outlets, a comparison of market shares within each case can guide equivalent source selection (Rössler, 2012). Suppose document samples are retrieved with search strings in different languages for different cases. In that case, these retrieved documents should be compared with human-annotated data to assess recall and precision, and misclassifications should be inspected (Mahl et al., 2022). A good example of such careful practice is provided by Rauh and De Wilde (2018), who offer a transparent account of how they evaluated and ensured cross-lingual equivalence in their selection of EU legislative debates.

3.2 Input Data Validation

3.2.1 Objective

After selecting the corpus for each language, researchers need to ensure that the produced units and textual representations are equivalent across cases.

[5] Biases may even arise with seemingly trivial sampling criteria - for instance, the keyword "Trump" may work reasonably well in non-US media for identifying coverage of the former US president, but may generate precision problems in US media (which may also cover other members of the family to some extent), and it will completely miss relevant transcripts of "The Late Show", since its host Colbert systematically replaces that name with slanderous nicknames.

In most comparative research, this may require actions such as harmonising unitization or stripping uninformative, idiosyncratic content across cases. When dealing with multilingual text, linguistic perspectives also come into play. Factors like how languages segment sentences into words, compound word formation, function word affixation, and morphological variations demand attention. Researchers are encouraged to reflect on and empirically test the equivalence of the data inputs across languages and cases. Importantly, the input data should be selected with the intended analytical model in mind, as different approaches (e.g., dictionary methods, topic models, or transformers) may place different demands on the structure and quality of input data.

3.2.2 Considerations

Data input validation first concerns unitization, i.e., the determination of what level of text relevant meanings are expressed and to what units of text they pertain. For multilingual analyses, this is especially important as differences in the meaning contained within each unit of text may confound equivalent measurement. Cultural, stylistic, and linguistic differences affect unitization, such as in the case of languages that prefer long (e.g., German) or short (e.g., English) sentences, or in Chinese speech, which does not segment words in writing, resulting in substantively different meanings of paragraphs. To ensure comparability, researchers may need to adapt the unit of analysis per language—for instance, combining several English sentences to match a dense German one, or using word segmentation in Chinese. Justifying these choices with linguistic and contextual insight is key to valid cross-language analysis.

In CTA, unitization does not end here, however; rather, texts need to be further broken down into so-called “tokens” - typically derived from words - whose distribution over different documents presents the primary source of information for many computational algorithms. However, tokenization is far from language-neutral. When deciding on the appropriate set of tokenisation steps, researchers need to consider in what form relevant meanings are encoded in the text. To ensure that equivalent information is retained in multiple languages and cases, it may be required to use different preprocessing strategies for different languages and cases so that the obtained textual representations are comparable and adequate for the intended analysis. For instance, many languages create numerous distinct tokens for the same word, depending on its case, gender, etc., which appear as independent from one another to a computer unless harmonised. In English, this may be

addressed by stemming (as used by Muddiman & Stroud, 2017). However, in languages that do not inflect words solely at the end, more sophisticated forms of lemmatization are required. For example, Hebrew has pre-, in-, and suffixes: "לִשְׂנֹא", "שְׂנֹאו", "שׂונא", "תִשנא" all reduce to the same root meaning "hate," requiring more advanced processing to treat them as equivalent.

Similarly, information expressed by conjunctions, prepositions, and other tokens typically discarded as "stopwords" in English (e.g., Muddiman & Stroud, 2017) is often conveyed through morphological variation, word order, or other mechanisms in other languages. For example, the English phrase "son of a bitch" risks losing key relational meaning if "of" is removed as a stopword. In contrast, the equivalent Arabic expression "اِبن العَاهِرَة" retains its structure and meaning through morphological encoding, illustrating how English-centric preprocessing can inadvertently distort multilingual semantic comparisons. These linguistic properties may confound a comparative study of uncivil language in unforeseen ways.

Additionally, in certain languages, incivility might manifest through the use of explicit swear words, while in others, it might be expressed through specific multi-word phrases. In Dutch, the intensifying use of disease-related terms such as *kankerlijer* (cancer sufferer) or *tyfushond* (typhus dog) creates compound insults that carry strong emotional weight, a pattern less common in English. Furthermore, the target of uncivil language is more easily observable in languages that hardwire grammatical function in particular word inflections (e.g., Turkish encodes the accusative case). Similarly, when one is interested in the extent to which uncivil language is gendered, it helps to consider how each of the languages under study distinguishes genders (if at all).

To determine input data validity, some extent of linguistic familiarity – and to the extent that pragmatic equivalence is required, cultural familiarity – is necessary. Researchers are generally advised to consult with native speakers and case experts when preprocessing unfamiliar languages and carefully validate outputs from standard pre-processing tools (such as those in NLTK, Bird et al., 2009), which may not generalize across languages. Tasks such as token reduction[6]. and word sense disambiguation[7] must be adapted

[6]For example, synthetic languages (that express grammar by means of affixes or morphology) contain significantly fewer highly common words than English, such that discounting the most common terms mostly eliminates articles, prepositions, conjunctions etc. in English, but reaches well into common nouns used for construct formation in other languages (e.g., בית [house] in Hebrew, which appears as part of very many construct state nouns: בית ספר [school], בית משפט [court], בית כנסת [synagogue])

[7]For example, disambiguation may be worth considering for important homonyms, where

with care, and multilingual preprocessing should be approached not as a standardized pipeline, but as a language-sensitive design challenge.

3.2.3 Evaluation

Ideally, researchers can offer evidence that input data alignment is established. If an analysis relies on preprocessing, a way to test the validity of the input data is to check if the meaning that is studied can still be easily recognised from the preprocessed text. At the same time, all variations that are lost should be immaterial to the pursued research question. One strategy that should help with detecting validity issues and biases in multilingual analysis at the level of preprocessing is to translate a few texts before preprocessing them and subject them to the respective preprocessing stages. The key test of cross-lingual validity is that the representations obtained after preprocessing map well onto one another, such that how relevant information is expressed (e.g., by single tokens, by token collocation patterns, or by sequentially arranged tokens) is similar between languages and cases. In studies where translation is part of the preprocessing stage, such as multilingual CTA projects that bring corpora into a shared representational space (e.g., English language text), back-translation is an effective validation technique. By iteratively applying machine translation tools, this method assesses the semantic stability of translated texts by testing whether repeated cycles of translation yield consistent and interpretable results (Chew et al., 2025). As one example of a study that presents extensive input validation steps, Segev (2019) discusses and argues for the selection of certain word types and the exclusion of others to increase the cross-linguistic and cross-national validity of the measurement instrument.

3.3 Process Validation

3.3.1 Objective

In addition to striving for equivalent textual representations, the researcher must also ensure the equivalence of the intended modeling strategy across languages and cases. In any use of CTA, the primary consideration governing validity is whether the algorithm indeed recognises the meaning intended to be measured. In multilingual projects, researchers must confirm that

the same spelling is used to express different meanings (e.g., 'rock', which can be a mass of stone, a musical genre, a movement, and more; some languages contain many more homonyms than others, so especially for Semitic languages, Part Of Speech tagging may be necessary to efficiently distinguish the many possible meanings of identical character sequences).

the algorithms are equally effective in all languages and cases. In practical terms, validation efforts concern model development, model selection, and model calibration.

3.3.2 Considerations

For their recognition of incivility, Muddiman and Stroud (2017) relied on a carefully validated dictionary of one-word tokens. Transferring the same design to different languages, however, numerous sources of bias arise from linguistic and context-related factors that may compromise the effectiveness of the same modeling strategy. A deeper understanding of the algorithmic design is a prerequisite for understanding how modeling assumptions and linguistic features interact across languages.

Most rule-based algorithms (e.g., dictionaries or search strings) offer considerable researcher control and thus generally facilitate reflection upon modeling choices across languages. Enabled to recognize that terms included in Muddiman and Stroud's (2017) incivility dictionary are polysemic, context-dependent, or require multi-word expressions in other languages, researchers can relatively easily add disambiguation criteria or shift toward a multi-word enabled strategy, while adjusting keywords for case-specific contexts as needed. For example, in Quebecois French, liturgical terms like *tabarnak* and *câlice* function as strong profanities and would likely be included in a dictionary of incivility for that context. In contrast, the same terms would appear semantically benign in continental French and thus be excluded. This contrast underscores how regional cultural variation affects which expressions register as uncivil and must be accounted for in multilingual modeling.

Matters are more complicated for unsupervised, supervised algorithms, or more recent textual representations such as embeddings, where association measures are typically hard-coded into the algorithm. Researchers must assess whether document-level word co-occurrences (common in classic machine learning) suffice as representations of the expressed meaning, or whether more restrictive measures need to be obtained (e.g., by designing specific features that focus estimation on particular patterns in the data). Moreover, unsupervised algorithms tend to make strict assumptions about the generative structure of a text. For instance, many clustering algorithms assume that every token belongs to exactly one cluster, while factor analytic or topic modeling algorithms permit that the same token may contribute to more than one cluster. Especially in languages with many homonyms (i.e., words that have the same spelling) or words that appear as part of different

multi-word expressions, some clustering solutions are inappropriate for certain analyses, as the same token cannot be constrained to one role exactly [8]. Languages that form compound words (that is, two or more words joined together to create a single, distinct word with a unified meaning) may pose issues as key terms in the analysis are too infrequent to constitute recognised patterns. For instance, if we were to search for patterns in the topical structure of uncivil comments, the algorithm has much better chances to recognize *welfare* and *cheat* (both words common also in other contexts) as a pattern than *Sozialbetrüger* in German, a relatively rare term unique to the topic). Especially for supervised algorithms, providing substantive justifications for the choice of specific algorithms or hyperparameters often presents a major challenge, as very little is known how exactly modelling language in a hierarchical fashion (e.g., decision tree algorithms), in a vector space (e.g., SVM), or a neural network might affect classification. While it is evident (and plausible) that such choices matter, virtually no theory is available for justifying specific decisions.

Such considerations are not only relevant for models that rely on word co-occurrences (such as bag-of-words models), but also for models that represent words as embeddings. For example, word embeddings represent each word as a vector of real numbers that is based on the premise that words that appear a lot near each other in a set of documents should have similar vector representations (for an overview, see Rodriguez & Spirling, 2022). While early embeddings were primarily developed for English and faced challenges in other languages due to data scarcity and linguistic differences, recent advances have produced embeddings that cover dozens of languages (e.g., Polyglot) as well as multilingual embedding systems (De Vries, 2021; Licht, 2023), with promising cross-lingual alignment (Chew et al., 2025). Nevertheless, theoretical concerns persist regarding how grammatical word order and syntactic structures—such as the Subject-Verb-Object order in English versus Subject-Object-Verb or other variations in other languages—may influence the semantic quality and comparability of embeddings. Understanding these effects remains an important open question, especially when embeddings are used for measurement across languages.

Returning to our ongoing example of detecting uncivil language, different languages may have different swearing cultures, which will have implications for either approach. For instance, in some languages, name-calling

[8] For example, the English phrase *"parliamentary inquiry committee chairman"* consists of four tokens, each of which also co-occurs in many other contexts. By contrast, the equivalent German word *"Untersuchungsausschussvorsitzender"* is a single, very rare token that does not co-occur with others in the same way.

relies on well-known slurs or offensive adjectives that frequently co-occur and are easy to detect. In others, incivility may take more subtle or culturally specific forms. In Dutch, for instance, calling someone a *pannenkoek* (pancake) is mildly offensive. It is roughly akin to calling someone an idiot, even though the literal meaning is benign. In some online subcultures, such as incel communities, incivility involves the creative coining of novel terms (*neologisms*) or the reappropriation of innocuous words (e.g., *foid*, *roastie*, or *skype* used derogatorily), which can be difficult to detect without cultural or contextual knowledge. In East Asian languages like Korean, supervised machine learning may, to a significant extent, simply rely on honorific forms to decide that no incivility can be present. In contrast, in most Western languages, such a form is unavailable or degraded, and thus unavailable for classification. When measuring uncivil language, such cultural differences should be explicitly acknowledged in the codebook to be used for annotating training documents for multilingual supervised classification or in the interpretation of topics in a multilingual topic model.

3.3.3 Evaluation

The easiest way to ascertain process validity is to subject a few texts to the algorithm's transformation and determine whether the represented information (e.g., associations) reflects meaningful features of the processed text, in the case of multilingual analysis, whether it reflects the same meaningful features in equivalent ways. In general, for multilingual analysis, it is desirable to determine that equivalent modeling choices indeed emphasise and discount equivalent information. If researchers create their own set of manually annotated training documents that inform the model, testing intercoder-reliability across languages and cases for these training materials serves as another process validation strategy. Calculating metrics like Machine Translation Accuracy (MTA) can serve as an effective validation strategy for comparing the multilingual outputs of various models. For an implementation that assesses the consistency of top words per topic across languages via MTA, refer to Lind et al. (2022). Validation strategies encompass further activities like visualising how inputs relate to outputs and identifying features that are most influential in the model's decision-making process (see, for example, the LIME technique as implemented in Ho and Chan (2023)). When relying on generative LLMs like ChatGPT, researchers can, for example, interrogate these models to explain their reasoning for why they annotate text in one way or another, and they can do so in multiple languages (see e.g., Kuzman et al., 2023). As a positive example, Maurer

and Diehl (2020) offer important details on how they validated their French and English dictionaries against crowd-sourced benchmarks and by further employing a qualitative pre-analysis of individual keywords.

3.4 Output Validation

3.4.1 Objective

The fourth validation stage concerns the quality of the obtained measures and their equivalence across languages and across cases concerning criterion validity (the obtained measures correlate with a trusted third variable) and face validity (the results appear plausible for each case) (Krippendorff, 2004). For comparative multilingual applications, output validation needs to be considered for each included language and case.

3.4.2 Considerations

Our discussion on output validation focuses on the type and quality of a third variable that can serve as a trusted benchmark for case and language comparisons. Researchers are advised to assess available third variables for either the same construct (for testing convergent validity) or distinctly different constructs (for testing divergent validity). When selecting or developing a human-coded benchmark as a third variable, as was the case in Muddiman and Stroud's (2017) study, its quality can be assessed by inspecting the quality of the codebooks, human coder training, and intercoder reliability tests. Codebook definitions should apply to all languages and cases (i.e., etic constructs), while rules and examples may need to be adjusted to different contexts and settings (e.g., racist incivility tends to focus on different ethnic groups in different countries). As with monolingual projects, it is advantageous to train all involved coders in joint sessions and to clarify issues or adjust the codebook collaboratively (Rössler, 2012). Intercoder reliability tests should cover reliability across languages/cases as well as within languages/cases (Peter & Lauf, 2002). If all coders are skilled in one language, they can code the same material to establish intercoder reliability (Hopmann et al., 2016). Care should be taken to ensure that the manually coded material is representative of all languages and cases. To achieve this, a random subset per language and case can be translated into a common language. To further assess intercoder reliability within language/case, at least two coders code the original language material per case. If a measured meaning is far more dominant, or largely absent, in one compared case, careful interpretation of performance scores is crucial; high accuracy does

not necessarily imply valid discrimination (e.g., if incivility appears in only 3% of cases, not coding it achieves 97% accuracy). Employing chance- and category frequency-corrected metrics might be necessary for informative measurement validity. For example, when examining uncivil language, do we see a higher recall or precision in some languages than others? And what features are predictive of observed differences?

3.4.3 Evaluation

Applying the convergent validation strategies involves the selection of a high-quality gold standard as described above and the calculation of recall and precision scores for each language and case in the dataset. For example, Lind et al. (2021) compare the results from a supervised classification approach with a benchmark coded by native speakers. Temporão et al. (2018) assess the convergent and predictive validity of their computationally obtained ideological scaling measures of social media users. Another key strategy for output validation, then, is to examine misclassifications carefully (Ho & Chan, 2023). Systematic issues in precision are detected by inspecting positively classified but irrelevant documents. Recall issues can be examined by drawing a random sample of texts for manual classification and examining which texts are missed by the model. When a proper gold standard exists, examining false positives and negatives can uncover possible biases and direct enhanced performance of the employed tools. Last but not least, it is of course advisable to critically check the results manually with case experts for their plausibility.

4 Conclusion

As the community of social scientists who study various languages at scale continues to globalize and the comparative turn in computational text analysis requires theoretical grounding, we have provided a framework for validating social science measurements across languages and cases.

Following a definition of validity in the context of such projects, the framework includes practical recommendations for assessing the validity of etic constructs measured across diverse languages and cases. It covers techniques for validating equivalent data sources and sample selection, preparation of equivalent input features, selection of equivalence data processing methods, attainment of equivalent measurements, and ultimately achieving equivalent mapping and thus comparable results. Despite advancements in CTA, collaboration with experts in relevant languages and regions, as well as knowledge about training procedures, remains crucial.

Issues around resource and retrieval bias have not disappeared in the age of LLMs, which, for many social science applications, need to be fine-tuned to identify a particular construct in smaller language-specific samples of data.

One immediate question that arises is how to deal with measures when validation techniques raise validity concerns that cannot be resolved? A first and crucial mitigation action is to report and reflect on the detected problems, as it enables future research to build on better information on related measurements. This way, the field can accumulate knowledge, build theory, and develop further recommendable strategies. The second is to make a considered decision about the extent to which the measurements can be used to make substantive comparative statements about the cases. The measurements may not be suitable for comparative statements, but they may be suitable for statements per case or for comparisons among a subset of the cases. Furthermore, researchers can explore error correction methods to account for misclassifications (i.e., Bachl & Scharkow, 2017; TeBlunthuis et al., 2023).

In light of this need to attend systematically to issues of validation in CTA, questions about the practical implementation of such validation immediately arise. Executing the validation requirements outlined in the framework requires significant resources. However, since the quality of the analysis is crucially dependent on the quality of the data and measurements, these are resources well-spent. Many of these steps can be supported by better research infrastructures (e.g., the Masakhane project, which works on African languages, Nekoto et al., 2020, or OPTED) and open science initiatives. For instance, curated inventories for data sources and tools that support the input data and process stages can save research teams valuable time and resources. These services can also facilitate collaboration between research partners and provide access to language and case experts from diverse backgrounds, which is essential for all validation stages (see Spirling, 2023, for a similar argument in the context of LLMs).

The framework we propose outlines an ideal standard for validation in comparative text analysis, especially for projects that aim for conceptual and measurement equivalence across languages and cases. While it sets out what best practice looks like, we acknowledge that fully adhering to this standard may not always be feasible in large-scale studies. Even if resources were to improve significantly in the coming years and a project had a very large budget, the kind of methodological rigor and emphasis on human evaluation by native speakers and case experts that are central to our framework mean that projects covering 50+ countries and languages are

hardly recommendable. For projects that do aim at such scale, we recommend prioritizing validation for a strategically selected subset of cases (e.g., high-variance contexts or linguistic clusters), using these as benchmarks for broader interpretation. While such strategies cannot fully replace comprehensive validation, they can help identify systematic issues and improve the interpretability of large-scale comparisons.

A limitation of the framework is that it does not provide a detailed breakdown of the input, process, and output validation stages for specific analytical methods, such as distinct preprocessing steps or particular approaches like dictionary methods, topic models, or transformer-based architectures. As the framework is intended to be broadly applicable across methods, such detail was beyond the scope of this article. Nonetheless, we hope it offers a useful structure to support more specific extensions in future work.

In conclusion, we encourage future research to define constructs comprehensively; to document and justify methodological choices; to report any validation efforts taken; and to prudently consider the detected validity concerns in the subsequent use of measurements. Not only should authors dedicate some space to enabling readers to follow taken validation steps and thus build confidence in the measurement, but reviewers need to demand and scrutinise presented validation efforts, and editors need to set aside the requisite space for the purpose. Transparency not only facilitates an informed assessment of measurement validity in CTA but also lays the groundwork for systematic theoretical and methodological discussions on best practices in validation.

5 Funding Note

This work was supported by funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 951832.

6 Acknowledgement

We sincerely thank the reviewers for their valuable feedback and thoughtful suggestions.

References

Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review, 95*(3), 529–546. https://doi.org/10.1017/S0003055401003100

Bachl, M., & Scharkow, M. (2017). Correcting measurement error in content analysis. *Communication Methods and Measures*, *11*(2), 87–104. https://doi.org/10.1080/19312458.2017.1305137

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2022). Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, *16*(1), 1–18. https://doi.org/10.1080/19312458.2021.1916027

Baden, C., Dolinsky, A., Lind, F., Pipal, C., Schoonvelde, M., Shababo, G., & van der Velden, M. A. (2022). Integrated standards and context-sensitive recommendations for the validation of multilingual computational text analysis. *Horizon Project OPTED: Deliverable 6.2*.

Baden, C., Jungblut, M., Micevski, I., Stalpouskaya, K., Tenenboim-Weinblatt, K., Berganza Conde, R., Dimitrakopoulou, D., et al. (2018). The infocore dictionary: A multilingual dictionary for automatically analyzing conflict-related discourse.

Baum, M. A., & Zhukov, Y. M. (2019). Media ownership and news coverage of international conflict. *Political Communication*, *36*(1), 36–63. https://doi.org/10.1080/10584609.2018.1493002

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bird, S., Loper, E., & Klein, E. (2009). *Natural language processing with python*. O'Reilly Media Inc.

Birkenmaier, L., Lechner, C. M., & Wagner, C. (2024). The search for solid ground in text as data: A systematic review of validation practices and practical recommendations for validation. *Communication Methods and Measures*, *18*(3). https://doi.org/10.1080/19312458.2023.2285765

Chan, C. H., Zeng, J., Wessler, H., Jungblut, M., Welbers, K., Bajjalieh, J. W., & Althaus, S. L. (2020). Reproducible extraction of cross-lingual topics (rectr). *Communication Methods and Measures*, *14*(4), 285–305. https://doi.org/10.1080/19312458.2020.1812559

Chew, E., Chakraborti, M., Weisman, W., & Frey, S. (2025). Evaluating machine translation solutions for accessible multi-language text analysis: A back-translation based approach. *Computational Communication Research*, *7*(1), 1–24. https://doi.org/10.5117/CCR2025.1.001.CHEW

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

De Vries, E. (2021). The sentiment is in the details: A language-agnostic approach to dictionary expansion and sentence-level sentiment analysis in news media. *Computational Communication Research*, *4*(2), 424–462. https://doi.org/10.5117/CCR2021.2.006.DEVR

Demeter, M. (2019). The winner takes it all: International inequality in communication and media studies today. *Journalism & Mass Communication Quarterly*, *96*(1), 37–59. https://doi.org/10.1177/1077699018804506

Dolinsky, A. O., Schoonvelde, M., Pipal, C., Baden, C., Lind, F., Shababo, G., Van der Velden, M., & Zalik, A. (2024). Challenges for multilingual computational text analysis researchers: Evidence from a survey of social scientists. https://osf.io/9mybf_v2

Esser, F., & Vliegenthart, R. (2017). Comparative research methods. In *The international encyclopedia of communication research methods* (pp. 1–22). Wiley. https://doi.org/10.1002/9781118901731.iecrm0035

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press. https://doi.org/10.23943/princeton/9780691207551.001.0001

He, J., & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture, 2*(2), 2307–0919. https://doi.org/10.9707/2307-0919.1111

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not weird. *Nature*, *466*(7302), 29–29. https://doi.org/10.1038/466029a

Henriksen, D. (2016). The rise in co-authorship in the social sciences (1980–2013). *Scientometrics*, *107*(2), 455–476. https://doi.org/10.1007/s11192-016-1849-x

Ho, J. C. T., & Chan, C. H. (2023). Evaluating transferability in multilingual text analyses. *Computational Communication Research*, *5*(2). https://doi.org/10.5117/CCR2023.2.003.HO

Hopmann, D. N., Esser, F., de Vreese, C. H., Aalberg, T., van Aelst, P., Berganza, R., & Strömbäck, J. (2016). How we did it: Approach and methods. In C. H. de Vreese, F. Esser, & D. N. Hopmann (Eds.), *Comparing political journalism* (pp. 10–21). Routledge. https://doi.org/10.4324/9781315622286-2

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, *15*(8), e12432. https://doi.org/10.1111/lnc3.12432

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Sage Publications.

Kuzman, T., Mozetič, I., & Ljubešić, N. (2023). Chatgpt: Beginning of an end of manual linguistic data annotation? use case of automatic genre identification.

Laurer, M. (2023). Lowering the language knowledge barrier – investigating deep transfer learning and machine translation for multilingual analyses of political texts. *Computational Communication Research*, *5*(2), 1–28. https://doi.org/10.5117/CCR2023.2.001.LAUR

Licht, H. (2023). Cross-lingual classification of political texts using multilingual sentence embeddings. *Political Analysis*, *31*(3), 366–379. https://doi.org/10.1017/pan.2022.29

Licht, H., & Lind, F. (2023). Going cross-lingual: A guide to multilingual text analysis. *Computational Communication Research*, *5*(2), 1–31. https://doi.org/doi.org/10.5117/CCR2023.2.2.LICH

Lind, F., Eberl, J.-M., Eisele, O., Heidenreich, T., Galyga, S., & Boomgaarden, H. G. (2022). Building the bridge: Topic modeling for comparative research. *Communication Methods and Measures*, *16*(2), 93–118. https://doi.org/10.1080/19312458.2021.2015572

Lind, F., Heidenreich, T., Kralj, C., & Boomgaarden, H. G. (2021). Greasing the wheels for comparative communication research: Supervised text classification for multilingual corpora. *Computational Communication Research*, *3*(3), 1–30. https://doi.org/10.5117/CCR2021.3.001.LIND

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, *23*(2), 254–277. https://doi.org/10.1093/pan/mpu019

Mahl, D., von Nordheim, G., & Guenther, L. (2022). Noise pollution: A multi-step approach to assessing the consequences of (not) validating search terms on automated content analyses. *Digital Journalism*, *11*(2), 298–320. https://doi.org/10.1080/21670811.2021.2023931

Mate, A., Sebök, M., Wordliczek, L., Stolicki, D., & Feldmann, Á. (2023). Machine translation as an underrated ingredient? solving classification tasks with large language models for comparative research. *Computational Communication Research*, *5*(2), 1–34. https://doi.org/10.5117/CCR2023.2.004.MATE

Maurer, P., & Diehl, T. (2020). What kind of populism? tone and targets in the twitter discourse of french and american presidential candidates. *European Journal of Communication*, *35*(5), 453–468. https://doi.org/10.1177/0267323120940903

Muddiman, A., & Stroud, N. J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication*, *67*(4), 586–609. https://doi.org/10.1111/jcom.12313

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., & Bashir, A. (2020). Participatory research for low-resourced machine translation: A case study in african languages. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2144–2160. https://doi.org/10.18653/v1/2020.findings-emnlp.195

O'Grady, W., Archibald, J., Aronoff, M., & Rees-Miller, J. (2017). *Contemporary linguistics: An introduction* (7th). Bedford/St. Martin's.

Palicki, S., Walter, S., van Atteveldt, W., Beazer, A., & Bravo, I. (2023). Selecting relevant documents for multilingual content analysis: An evaluation of keyword and semantic similarity search approaches. *Computational Communication Research*, *5*(2), 1–54. https://doi.org/10.5117/CCR2023.2.PALI

Peter, J., & Lauf, E. (2002). Reliability in cross-national content analysis. *Journalism and Mass Communication Quarterly*, *79*(4), 815–832. https://doi.org/10.1177/107769900207900403

Proksch, S.-O., Lowe, W., Wäckerle, J., & Soroka, S. (2019). Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly*, *44*(1), 97–131. https://doi.org/10.1111/lsq.12212

Rathje, S., Mirea, D. M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). Gpt is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34), e2308950121. https://doi.org/10.1073/pnas.2308950121

Rauh, C., & De Wilde, P. (2018). The opposition deficit in eu accountability: Evidence from over 20 years of plenary debate in four member states. *European Journal of Political Research*, *57*(1), 194–216. https://doi.org/10.1111/1475-6765.12203

Reber, U. (2019). Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora. *Communication Methods and Measures*, *13*(2), 102–125. https://doi.org/10.1080/19312458.2019.1610977

Rinke, E. M., Dobbrick, T., Löb, C., Zirn, C., & Wessler, H. (2022). Expert-informed topic models for document set discovery. *Communication Methods and Measures*, *16*(1), 39–58. https://doi.org/10.1080/19312458.2021.2014861

Rodriguez, P. L., & Spirling, A. (2022). Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics*, *84*(1), 101–115. https://doi.org/10.1086/714944

Rössler, P. (2012). Comparative content analysis. In F. Esser & T. Hanitzsch (Eds.), *The handbook of comparative communication research* (pp. 481–490). Routledge.

Salganik, M. J. (2017). *Bit by bit: Social research in the digital age*. Princeton University Press.

Scharkow, M., & Trepte, S. (2023). National diversity at conferences of the international communication association [Published online]. *Annals of the International Communication Association*. https://doi.org/10.1080/23808985.2023.2175968

Segev, E. (2019). From where does the world look flatter? a comparative analysis of foreign coverage in world news. *Journalism*, *20*(7), 924–942. https://doi.org/10.1177/1464884917709344

Shababo, G., Baden, C., Dolinsky, A., Lind, F., Pipal, C., Schoonvelde, M., Smoliarova, A., van der Velden, M. A. C. G., & Zalik, A. (2023). How do linguistic differences impact computational text analysis methods? a road map for future validation and integrated strategy development [OPTED Deliverable 6.3]. https://opted.eu/fileadmin/user_upload/k_opted/OPTED_Deliverable_D6.3.pdf

Sigismondi, P. (2018). Exploring translation gaps: The untranslatability and global diffusion of 'cool'. *Communication Theory*, *28*(3), 292–310. https://doi.org/10.1111/comt.12123

Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., & Boomgaarden, H. G. (2020). In validations we trust? the impact of imperfect human annotations as a gold standard on the quality of validation of automated content

analysis. *Political Communication*, *37*(4), 550–572. https://doi.org/10.1080/10584609.2020.1811890

Spirling, A. (2023). Why open-source generative ai models are an ethical way forward for science. *Nature*, *616*(7957), 413–413. https://doi.org/10.1038/d41586-023-02395-7

Stecker, M., Balluff, P., Lind, F., Dinhopl, C., Waldherr, A., & Boomgaarden, H. G. (2024). Tools of the trade–when are software tools mentioned in computational text analysis research? *Computational Communication Research*, *6*(1), 1–21. https://doi.org/10.55563/ccc/abcd12

TeBlunthuis, N., Hase, V., & Chan, C.-H. (2023). Misclassification in automated content analysis causes bias in regression. can we fix it? yes, we can! [arXiv preprint arXiv:2307.06483]. https://arxiv.org/abs/2307.06483

Temporão, M., Kerckhove, C. V., van der Linden, C., Dufresne, Y., & Hendrickx, J. M. (2018). Ideological scaling of social media users: A dynamic lexicon approach. *Political Analysis*, *26*(4), 457–473. https://doi.org/10.1017/pan.2018.19

van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2-3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

Volk, S. C. (2022). *Comparative communication research*. Springer Fachmedien Wiesbaden.

Wirth, W., & Kolb, S. (2004). Designs and methods of comparative political communication research. In F. Esser & B. Pfetsch (Eds.), *Comparing political communication: Theories, cases, and challenges* (pp. 87–111). Cambridge University Press.

Wozniak, A., Wessler, H., Chan, C. H., & Lueck, J. (2021). The event-centered nature of global public spheres: The un climate change conferences, fridays for future, and the (limited) transnationalization of media debates. *International Journal of Communication*, *15*, 27.